LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-222567

Exploration of methods to identify polymorphisms associated with variation in DNA repair capacity phenotypes

I. M. Jones, C. B. Thomas, T. Xi, H. W. Mohrenweiser, D. O. Nelson

July 5, 2006

Mutation Research

Exploration of methods to identify polymorphisms associated with variation in DNA repair capacity phenotypes

Irene M. Jones[1], Cynthia B. Thomas[1], Tina Xi[1], Harvey W. Mohrenweiser[2], David O. Nelson[1]

To be submitted to Mutation Research
July 2006

[1] Lawrence Livermore National Laboratory, Livermore, CA USA
[2] Oregon Health & Science University, Portland, OR USA

Corresponding author:
Irene M. Jones
L-441, PO 808
7000 East Avenue
Lawrence Livermore National Laboratory
Livermore, CA 94550 USA

Phone: (925) 423-3626
Fax: (925) 424-3130
e-mail: jones20@llnl.gov

Abstract
Elucidating the relationship between polymorphic sequences and risk of common disease is a challenge. For example, although it is clear that variation in DNA repair genes is associated with familial cancer, aging and neurological disease, progress toward identifying polymorphisms associated with elevated risk of sporadic disease has been slow. This is partly due to the complexity of the genetic variation, the existence of large numbers of mostly low frequency variants and the contribution of many genes to variation in susceptibility. There has been limited development of methods to find associations between genotypes having many polymorphisms and pathway function or health outcome. We have explored several statistical methods for identifying polymorphisms associated with variation in DNA repair phenotypes. The model system used was 80 cell lines that had been resequenced to identify variation; 191 single nucleotide substitution polymorphisms (SNPs) are included, of which 172 are in 31 base excision repair pathway genes, 19 in 5 anti-oxidation genes, and DNA repair phenotypes based on single strand breaks measured by the alkaline Comet assay. Univariate analyses were of limited value in identifying SNPs associated with phenotype variation. Of the multivariable model selection methods tested: the easiest that provided reduced error of prediction of phenotype was simple counting of the variant alleles predicted to encode proteins with reduced activity, which led to a genotype including 52 SNPs; the best and most parsimonious model was achieved using a two-step analysis without regard to potential functional relevance: first SNPs were ranked by importance determined by Random Forests Regression (RFR), followed by cross-validation in a second round of RFR modeling that included ever more SNPs in declining order of importance. With this approach 6 SNPs were found to minimize prediction error. The results should encourage research into utilization of multivariate analytical methods for epidemiological studies of the association of genetic variation in complex genotypes with risk of common diseases.

Key Words
Polymorphism, DNA repair, phenotype, genotype

Introduction

Elucidating the relationship between polymorphic sequences and risk of common disease is a challenge. For example, variation in the sequence of DNA repair genes among healthy people may contribute to risk of cancer, rate of aging, neurological disease and cardiovascular disease. Adding to the complexity, the impact of variation in exposures from diet, lifestyle, occupation and other environmental agents may be genotype dependent. As a consequence of the Human Genome Program, it has become feasible to systematically screen for DNA sequence variation in the human population, for example the HapMap project and the Environmental Genome Project. In recent years increasing technical ease and declining cost of genotyping has enabled and accompanied the increase in knowledge of genetic variation. Coupled with availability of large, well characterized cohorts, there are increasing opportunities to study the role of genetic variation in susceptibility to exposure and risk of disease. However, progress toward identifying the polymorphisms contributing to increased risk of common disease in the population (in contrast to family studies) has been slow. The nature of the genetic variation, with large numbers of mostly low frequency variants affecting both coding and noncoding sequences, has contributed to limited characterization of the functional impact of individual single nucleotide polymorphisms (SNPs) and, perhaps more important, the impact of this extensive variation on the activity of pathways in which multiple genes are required for full functionality. There has been limited research on methods to find associations between SNPs and function or health outcome that take existing genomic complexity into account.

With respect to cancer risk, utilization of functional assays providing quantitative measures of the ability of cells to repair DNA damage has been a powerful method for establishing the association of variation in DNA repair pathways with disease [1,2]. High heritability (0.65-0.80) for phenotypes associated with several repair pathways [3-5] provides evidence that genetic variation is a major contributor to variation in DNA repair phenotypes. The power of functional assays for measuring the activity of pathways resides in their ability to capture the impact of variation in genes/proteins in pathways without perfect knowledge of the variants or genes contributing to the variation in activity.

In contrast, genotyping/molecular epidemiology studies have had limited success in finding associations of individual polymorphisms with risk of common disease, including cancer [6]. For example, only six of 166 published molecular genetic association studies replicated three or more times yielded consistent results [7]. Issues of study design and data analysis are more likely responsible for these weak and inconsistent outcomes than lack of impact of genetic variation on susceptibility to disease.

In this study we present our exploration of several statistical methods for identifying polymorphisms associated with variation in DNA repair phenotypes. We used a candidate pathway approach to compare methods for assessing the association of variation of cellular DNA repair phenotypes in human cell lines with previously reported genotypes for 172 SNPs from a high proportion of base excision repair (BER) pathway genes (31 of 39) [8-10] and 19 SNPs among 5 anti-oxidation genes. The phenotypes studied include composite measures of single strand DNA breaks, abasic sites and alkali-labile sites present due to endogenous DNA metabolism, including but not limited to BER repair of oxidative and alkylation damage [11]

(referred to as "background"), and damage present due to the net effects of anti-oxidation and repair at two times after exposure to ionizing radiation, immediately after exposure and after a short period of repair.

Methods and Results

The sample set.
All genotype and phenotype analyses were performed in 80 lymphoblastoid cell lines of the DNA Polymorphism Discovery Resource (DPDR) [12], obtained from Coriell Institute of Medical Technology (Camden, NJ). These uses of these anonymous samples were reviewed and determined to be exempt from Human Subject requirements by the Lawrence Livermore National Laboratory (LLNL) Institutional Review Board.

Genotype dataset.
Resequencing methods to identify variation in DNA sequence have been described [13,14]. The genotype data used in this study were assembled from three sources: *http://greengenes.llnl.gov/dpublic/secure/reseq/* (LLNL studies) and *http://www.genome.washington.edu/projects/egpsnps/* and *http://egp.gs.washington.edu/*. SIFT (Sorting Intolerant From Tolerant) [15] scores for amino acid substitution variants were determined as described in [10]. Relevant to these analyses, 191 SNPs were identified in the 80 cell lines, 172 in the resequencing of 31 BER pathway genes and 19 in the screening of 5 anti-oxidation genes. All amino acid substitutions of frequency 0.01 or higher in the selected genes were included. In addition for exploratory purposes, 15 SNPs affecting the 5' UTR of 12 genes, allele frequency 0.10 or above and average frequency 0.22, were also included. Genotypes for 58 of the SNP loci were known in all 80 cell lines, whereas genotypes for 1-5 cell lines were missing for 103 SNPs and 35 were missing in 6-16 cell lines; the remaining 5 SNPs had been analyzed in only about half the cell lines. Missing genotype values for SNP loci were imputed to be wild type (i.e., AA). Many of the variant alleles exist at low frequency and thus are observed in only a small number of the cell lines studied. For example 92 of the 191 SNPs exist in only one of the 80 cell lines. Given that cross-validation procedures (see below) required that a SNP be present in at least two cell lines, the ability of this model selection technique to find SNPs of interest was limited to the 99 SNPs existing in at least two cell lines.

Phenotype dataset.
The alkaline Comet assay (modified from [16]) was used to measure endogenous DNA damage and the damage present as a function of time after exposure to ionizing radiation. An exponentially growing culture of each cell line was sampled once and viability determined by Trypan blue dye exclusion (average viability for the 80 cell lines was 90% at the time of analysis). Cells were sampled prior to radiation or exposed on ice to 5Gy (3 Gy/min) from a ^{137}Cs source and then sampled after 0 and 15 minutes incubation at 37°C. Slides for Comet analysis were prepared, placed in lysis buffer (1% Triton X-100, 10% DMSO, 89% stock lysing solution: 2.8M NaCl, 0.1M Na_2EDTA, 0.01M Trizma Base) at least overnight, then rinsed 3 x 10 min in 0.4M Tris, pH 7.5. Slides were covered with a fresh solution of 300 mM NaOH, 1 mM EDTA, final pH >13.0 for 60 min, then electrophoresed at 0.92 V/cm with current adjusted to 300 mamps for 25 min. Each slide was stained with 100 µl of ethidium bromide (2 µg/ml). Images of 50 cells on each of 2 slides were captured on a Zeiss Axioplan epi-illumination fluorescence

microscope (C. Zeiss, Oberkpchen, Germany) fitted with a CCD camera. Comet parameters were determined using Komet4.0©: Image Analysis and Data Capture (Kinetic Imaging, Ltd., Merseyside, England). The specific phenotype outcomes analyzed are listed under univariate analyses (below). The selection of Comet parameters for analysis was based on their prevalence in the literature (percent of DNA in tail and tail length) and on our determination that the Comet Distributed Moment (CDM) [17] parameter provided the best signal to noise ratio amongst matched slides of cells exposed to 5Gy that were electophoresed at the same time (data not presented).

Statistical analyses.

Univariate Analyses

For each SNP and six phenotype outcomes, we used an F-test from an analysis of variance to assess the difference in mean outcome among genotypes for that SNP. To control for multiple testing, the p-values were adjusted to control for False Discovery Rate using the procedure outlined by Benjamini and Hochberg [18]. The six different phenotype outcomes analyzed were: (1) Background levels (no radiation exposure), as measured by CDM; (2) amount of damage induced by 5 Gy, as measured by CDM immediately after irradiation on ice; (3) amount of damage repaired in 15 minutes, as measured by CDM; (4) percent of damage repaired, as measured by CDM (we transformed this variable by taking the square root of the percentage); (5) background levels, as measured by Tail Length; (6) background levels, as measured by percent of DNA in Tail (tail%DNA). The histograms of the distribution of these six CDM phenotype outcomes among the 80 cell lines are shown in Fig. 1.

None of the first three outcomes resulted in any SNP with an adjusted p-value <0.20. However, Percent Repaired (CDM) reported four SNPs with an adjusted p-value <0.20: one each in LIG1, MPG, NTHL1, and POLE, with adjusted p-values of 0.01, 0.14, 0.14, and 0.14 respectively. Background levels (Tail length) produced two SNPs in GPX1, both with an adjusted p-value of 0.16. Finally Background levels (TDNA) produced one SNP in RFC1, with an adjusted p-value of 0.11. Except for the GPX1 SNPs, all the above SNPs have very low allele frequency, and the effect was due to one (or at most two) cell lines with Aa (heterozygote) genotypes. Given the limited significance of these results, the detailed identities of these SNPs are not provided.

Multivariable Model Selection employing SIFT scores

Our first approach to multivariable model selection was based on SIFT scores. SIFT scores have been reported as being able to predict amino acid changes that affect function [15]. When using SIFT scores, the SNPs were added to a model in order of increasing SIFT score. Each time a set of SNPs with a given score was added to the model, the expected prediction error of this augmented model was re-calculated. This model selection process results in a set of nested models, with one model for each unique SIFT score. The "best" model is the one with the lowest expected prediction error. We tried two types of models for our SIFT-based model selection. The first model type, which we term "sum the alleles", creates a variable for each cell line consisting of the number of variant alleles found in that cell line across all the SNPs in the model. The value of the sum will depend on which SNPs are in the model, the number of which increases as we increase the SIFT cutoff threshold. We used leave-one-out cross-validation to estimate the expected prediction error for each "sum-the-alleles" model. The second model type is one based

on Random Forests Regression (RFR) [19]. For each set of SNPs, we created a RFR model and used the estimate of expected prediction error generated automatically by the random forests algorithm. Fig. 2 shows the results of SIFT-based model selection for the "sum the alleles" model type, using CDM as the outcome measure for four phenotypes. We see that background, damage induced and, marginally, percent damage repaired, produce models which decrease the expected prediction error for some set of SNPs with a low SIFT score. Fig. 3 compares the results of model selection via sum-the-alleles and random forests for the background CDM phenotype. We see that using only SIFT scores less than around 0.1 produces the best performing model of this type. We also see that the simplest model, just counting the number of variant alleles with low SIFT scores, does as well as Random Forests. The conclusion: using Random Forests out of the box provides no improvement over a very simple model.

Multivariable Model Selection ignoring SIFT scores

Because of the granularity of SIFT scores, the smallest non-null model in the above analysis contained the fifty-two SNPs with a SIFT score of 0. Next we tested whether a more complex model selection procedure that ignores SIFT scores might generate a parsimonious model. The "best" model for each of k=0,1,2,3,…,15 variables was separately estimated, where the "best" model is the one with the lowest expected prediction error. The model selected from the sixteen "best" models is the one with the lowest prediction error. We used importance measures from RFR to order the SNPs and select the top k SNPs. We then used leave-one-out cross-validation with another round of RFR, this time using only the k selected SNPs, to estimate the expected prediction error for a k variable model. Assessments of expected prediction error using leave-one-out cross validation requires that each SNP have a chance of appearing both in the training set and the test set. SNPs that had only one cell line with a non-wild genotype are not useful in assessing prediction error. Hence, we limited our analysis to SNPs with an non-wild genotype in at least two cell lines, reducing the SNPs to be examined to 99 SNPs. In addition, due to the large number of SNPs with only two non-wild genotypes, we extended the number of trees to be examined by random forests to 10,000.

A non-null model that reduced expected prediction error was detected only for one phenotype outcome, the background CDM outcome. Fig. 4 shows the relationship between number of variables and effect on expected prediction error. Table 1 shows the results of this analysis and selected attributes of the six SNPs that optimize prediction in this data set. Note: The two XRCC1 variant alleles are not in linkage disequilibrium.

Discussion

The results presented provide evidence that variation in quantitative DNA repair phenotypes is more effectively predicted by statistical approaches that incorporate multiple polymorphisms into the models than those that study one at a time. Of the methods explored, the best model in terms of reduction of expected prediction error required no prior knowledge of the 99 polymorphisms' potential impact on function. Although it was the most computationally demanding, this approach produced the most parsimonious model.

This exploratory study has many strengths and weaknesses. Among the strengths are that increasingly complex analytic methods are compared using the same data set for genotypes and phenotypes; a candidate pathway approach was utilized in which many SNPs from a large number the genes likely to be associated with the phenotypes were evaluated; amino acid

substitution SNPs were emphasized so that methods that predict functional impact might be used either in the analyses themselves or for assessing plausibility of results; the cell lines utilized have been resequenced for many other DNA repair genes and all cell lines and genotype data are in the public domain, such that these studies can readily be repeated and extended to other genes, pathways and phenotypes. The primary limitation of these studies is the small dataset of 80 cell lines. An additional limitation is occurrence of missing genotype data; our coding of missing data as wild type would tend to understate the effect of a variant.

In addition to an adequate sample size, success in detecting potential relationships between a phenotype and a large set of genotypes depends on having a reasonably close correspondence between the actual underlying relationship and the proposed statistical models that describe that relationship. It is usually assumed that the underlying genotype-phenotype relationship is one of a few genotypes having substantial effects, either independently or as a multigene interaction, and existing among a large number of genotypes having no effect at all. In this situation, approaches that search for a few big effects, either exhaustively [20] or heuristically [21], are usually effective. The difference between the two types of methods is one of scaleability: exhaustive methods run into a computational barrier after a couple of dozen SNPs, while heuristic methods such as the Random Forests approach used here scale well with the number of SNPs [22]. Results of modeling the base excision repair pathway by Sokhansanj and [23] are consistent with genetic variation leading to major reductions in activity of one protein or the simultaneous variation in several proteins being responsible for reduced BER pathway activity. In the later scenario, the underlying genotype-phenotype relationship is one in which the large number of SNPs existing in most individuals have a cumulative effect, but each effect is so small that detecting and quantifying it is nearly impossible, given the complexity and variability of the "background" genotype. In this situation, the above techniques will likely fail. Instead, much simpler models such as those which simply add up the number of variant genotypes may perform as well or better than more sophisticated approaches (see Millikan et al [24] where increasing number of variant alleles was associated with increasing risk of melanoma). Such approaches may perform even better still if one can provide a coarse screening of the SNPs to eliminate those which are likely *not* to have an effect on protein and/or pathway activity. Sorting by SIFT score is one relatively simple approach to screening, although the robustness of SIFT for identifying variants with only modest impact on activity is unknown. Although several approaches have been proposed for handling the first scenario above, methods to analyze data under the second scenario are much less well developed. One approach for variable selection that may work well under this scenario is a penalized likelihood approach like those used in "Least Angle Regression" methods [25, 26]

Given the preliminary nature of these studies, extensive discussion of the 6 SNPs identified in the 'best' model as associated with variation in background CDM is inappropriate. However, it is important to note several features of the results. With respect to plausibility, the 5 identified amino acid substitutions have low SIFT scores, consistent with likelihood of reduced function [15] and, interestingly, the 5'UTR identified is in a CpG island. Three of the SNPs have allele frequencies of ~0.03, hence their effect on phenotype implies codominance, and the potential that low frequency SNPs may be as influential at the population level as high frequency recessive SNPs. It is notable that in some cases the effect of some variants was to decrease the measured phenotype, while other variants were associated with decreased damage and/or increased repair. Detailed research into these effects, after validation in future studies, will provide insights into protein and pathway function.

In conclusion, this report illustrates the potential to relate variation in highly complex genotypes to variation in biological functions that may be relevant to human disease. Extension to more extensive genotypes, not limiting inclusion to presumptions about the role of a gene or impact of a SNP, and additional phenotypes, and research and development of multivariate model selection methods will lead to both improved predictive power and new insights into the underlying biology. The need to increase research on these and related data analysis and study design issues is enormous. The field of genetic epidemiology is poised to launch mega-projects in search of defining the genetic contribution to a number of common diseases using large, well characterized cohorts and low-cost, high throughput genotyping. The success of these efforts depends critically on research and development of study design and new approaches to analysis of the huge volume of data generated.

Acknowledgements
The authors dedicate this paper to Anthony V. Carrano and gratefully acknowledge the many influential roles that he played in their careers and in the development of genetic toxicology, human genomics and their interplay. We regret that the brevity of this report precluded citation of many related studies in the fields of genetic toxicology and human epidemiology. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. This research was funded in part by the Laboratory Directed Research and Development (LDRD) Program at LLNL. The LDRD Program is mandated by Congress to fund laboratory-initiated, long-term research and development (R&D) projects in support of the DOE and national laboratories' mission areas. The Director's Office LDRD Program at LLNL funds creative and innovative R&D to ensure the scientific vitality of the Laboratory in mission-related scientific disciplines.

References

1 M. Berwick and P. Vineis Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review, J. Natl. Cancer Inst. 92 (2000) 874-897.

2 M. Berwick, G. Matullo and P. Vineis Studies of DNA repair and human cancer: an update, in: S.H. Wilson and W.A. Suk (Eds.), Biomarkers of Environmentally Associated Disease: Technologies, Concepts and Perspectives, Lewis Publishers, Boca Raton, 2002, pp. 84-105.

3 J. Cloos, E.J.C. Nieuwenhuis, D.I. Boomsma, D.J. Kuik, M.L.T. van der Sterre, F. Arwert, G.B. Snow and B.J.M. Braakhuis Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripheral blood lymphocytes, J. Natl. Cancer Instit. 91 (1999) 1125-1130.

4 C.R. Kent, J.J. Eady, G.M. Ross and G.G. Steel The comet moment as a measure of DNA damage in the comet assay, Int J Radiat Biol 67 (1995) 655-660.

5 X. Wu, M.R. Spitz, C.I. Amos, J. Lin, L. Shao, J. Gu, M. de Andrade, N.L. Benowitz, P.G. Shields and G.E. Swan Mutagen sensitivity has high heritability: evidence from a twin study, Cancer Res 66 (2006) 5993-5996.

6 E.L. Goode, C.M. Ulrich and J.D. Potter Polymorphisms in DNA repair genes and associations with cancer risk, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1513-1530.

7 J.N. Hirschhorn, K. Lohmueller, E. Byrne and K. Hirschhorn A comprehensive review of genetic association studies, Genet Med 4 (2002) 45-61.

8 H.W. Mohrenweiser, D.M. Wilson, III and I.M. Jones Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes, Mutat. Res. 526 (2003) 93-125.

9 D.M. Wilson, 3rd and L.H. Thompson Life without DNA repair, Proc. Natl. Acad. Sci. USA 94 (1997) 12754-12757.

10 T. Xi, I.M. Jones and H.W. Mohrenweiser Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function, Genomics 83 (2004) 973-979.

11 T. Lindahl, P. Karran and R.D. Wood DNA excision repair pathways, Current Opin. in Genet. Dev. 7 (1997) 158-169.

12 F.S. Collins, L.D. Brooks and A. Chakravarti A DNA polymorphism discovery resource for research on human genetic variation, Genome Res. 8 (1998) 1229-1231.

13 H.W. Mohrenweiser, T. Xi, J. Vazquez-Matias and I.M. Jones Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1054-1064.

14 M.R. Shen, I.M. Jones and H. Mohrenweiser Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans, Cancer Res 58 (1998) 604-608.

15 P.C. Ng and S. Henikoff SIFT: Predicting amino acid changes that affect protein function, Nucleic Acids Res 31 (2003) 3812-3814.

16 N.P. Singh, M.T. McCoy, R.R. Tice and E.L. Schneider A simple technique for quantitation of low levels of DNA damage in individual cells, Exp. Cell. Res. 175 (1988) 184-191.

17 Y. Benjamini and Y. Hochberg Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological) 57 (1995) 289-300.
18 S.A. Roberts, A.R. Spreadborough, B. Bulman, J.B. Barber, D.G. Evans and D. Scott Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer? Am J Hum Genet 65 (1999) 784-794.
19 L. Breiman Random forests, Machine Learning 45 (2001) 5-32.
20 D.E. Comings, R. Gade-Andavolu, L.A. Cone, D. Muhleman, and J.P. MacMurray. A multigene test for the risk of sporadic breast carcinoma, Cancer 97 (2003) 2160-2170.
21 A. Bureau, J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, T.P. Keith and P. Van~Eerdewegh Identifying {SNPs} predictive of phenotype using random forests, Genet Epidemiol 28 (2005) 171-182.
22 J.H. Friedman Recent Advances in Predictive (Machine) Learning, PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology, Stanford Linear Accerator Center, Menlo Park, 2003.
23 B.A. Sokhansanj and D.M. Wilson, 3rd Estimating the effect of human base excision repair protein variants on the repair of oxidative DNA base damage, Cancer Epidemiol Biomarkers Prev 15 (2006) 1000-1008.
24 R.C. Millikan, A. Hummer, C. Begg, J. Player, A.R. de Cotret, S. Winkel, H. Mohrenweiser, N. Thomas, B. Armstrong, A. Kricker, L.D. Marrett, S.B. Gruber, H.A. Culver, R. Zanetti, R.P. Gallagher, T. Dwyer, T.R. Rebbeck, K. Busam, L. From, U. Mujumdar and M. Berwick Polymorphisms in nucleotide excision repair genes and risk of multiple primary melanoma: the Genes Environment and Melanoma Study, Carcinogenesis 27 (2006) 610-618.
25 B. Efron, T. Hastie, I.M. Johnstone and R. Tibshirani Least angle regression, Annals of Statistics 32 (2004) 407-499.
26 R. Tibshirani Regression shrinkage and selection via the lasso, J. Royal Statistical Society Series B 58 (1996) 267-288.

Table 1: Six SNPs were selected as associated with background single strand break DNA damage by multivariate model selection analysis ignoring SIFT scores. Damage was quantified by the Comet Distributed Moment (CDM), measured by the alkaline Comet assay. The 6 SNPs are presented in decreasing order of importance. Also presented are other aspects of the SNP: the relative importance of the SNP, its SIFT score, the distribution of genotypes among the 80 cell lines, the mean value of the phenotype for the AA genotype (homozygous wildtype), and the effect sizes (relative to AA) and standard errors for the other two genotypes. See Fig. 4 for the analysis that resulted in these six SNPs being selected.

SNP identity and attributes					Genotype			Phenotype *				
					Distribution among cell lines				Difference from AA and SE (Dif.)			
LLNL ID**	ID ****	effect	Rel. Imp.	SIFT	AA	Aa	aa	Mean AA	Aa Dif.	Aa Std. Err.	aa Dif.	aa Std. Err.
RFC1.1	rs3736168	5' UTR	2.3	-	56	14	10	23.4	-2.0	0.8	-0.8	0.9
RPA4.1***	rs2642219	33 Thr	1.3	0.06	43	18	19	23.1	0.4	0.7	-1.1	0.7
XRCC1.10	rs25487	399 Gln	1.3	0.02	46	30	4	23.1	-0.8	0.6	3.8	1.3
XRCC1.3	rs25496	72 Ala	0.7	0.00	75	5	0	22.8	2.6	1.2		
POLD1.1	rs3218773	19 His	0.3	0.00	76	4	0	23.1	-3.0	1.3		
LIG3.2	rs3136025	780 His	0.2	0.13	75	5	0	22.8	2.8	1.2		

* arbitrary units of CDM

** Gene name decimal LLNL SNP code

*** *RPA4* was previously denoted by the locus symbol *HSU24186*

**http://www.ncbi.nlm.nih.gov/SNP/index.html

Fig. 1. Histograms of the six DNA repair phenotype outcomes. Repair phenotypes are represented by damage measured by the alkaline Comet assay: (a) Background levels (no radiation exposure), as quantified Comet Distributed Moment (CDM); (b) amount of damage induced by 5 Gy immediately after irradiation on ice, as quantified by CDM; (c) amount of damage repaired in 15 minutes after irradiation, as quantified by CDM; (d) percent of damage repaired (Pct. Repaired) in 15 minutes, as quantified by CDM (we transformed this variable by taking the square root of the percentage); (e) background levels, as measured by Tail Length; (f) background levels, as quantified by percent of DNA in Tail (Tail%DNA).

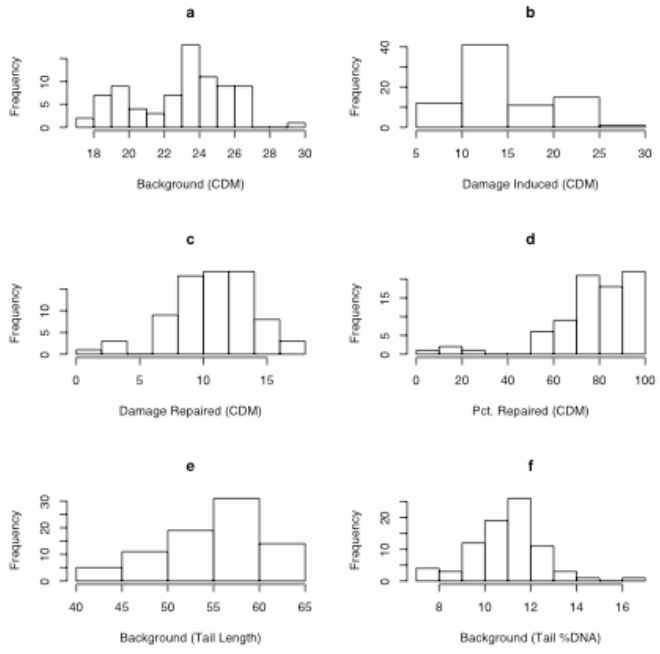

Fig. 2, Expected prediction error for predicting four phenotypes as a function of SIFT score cutoff for the simple model that predicts phenotype as a linear function of the sum of the number of variant alleles for all SNPs with a SIFT score not exceeding the SIFT cutoff. The horizontal line shows the expected prediction error for the null model, i.e., the variance of the phenotype. In all cases the phenotype is the CDM outcome measured by the alkaline Comet assay: (a) Background levels (no radiation exposure); (b) amount of damage induced by 5 Gy, measured immediately after irradiation on ice; (c) amount of damage repaired in 15 minutes; (d) percent of damage repaired in 15 minutes (we transformed this variable by taking the square root of the percentage). All 191 SNPs were used for these models.

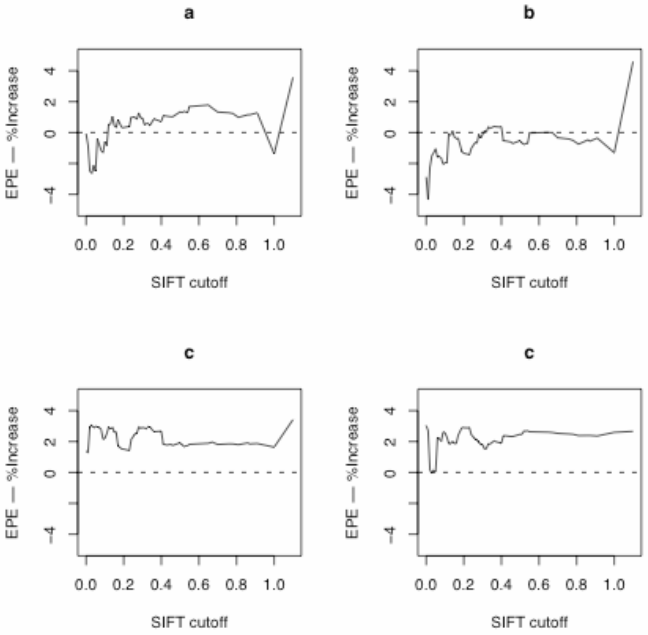

Fig. 3. Expected prediction error for predicting background DNA damage as a function of SIFT score cutoff for two model classes. The dashed line corresponds to a model that sums up the number of variant alleles for all SNPs with a SIFT score not exceeding the SIFT cutoff. The solid line corresponds to a Random Forests Regression model based on all SNPs with a SIFT score not exceeding the SIFT cutoff. The dashed horizontal line shows the expected prediction error for the null model, i.e., the variance of the background values as quantified by the Comet Distributed Moment in the alkaline Comet assay. All 191 SNPs were used for both of these models.

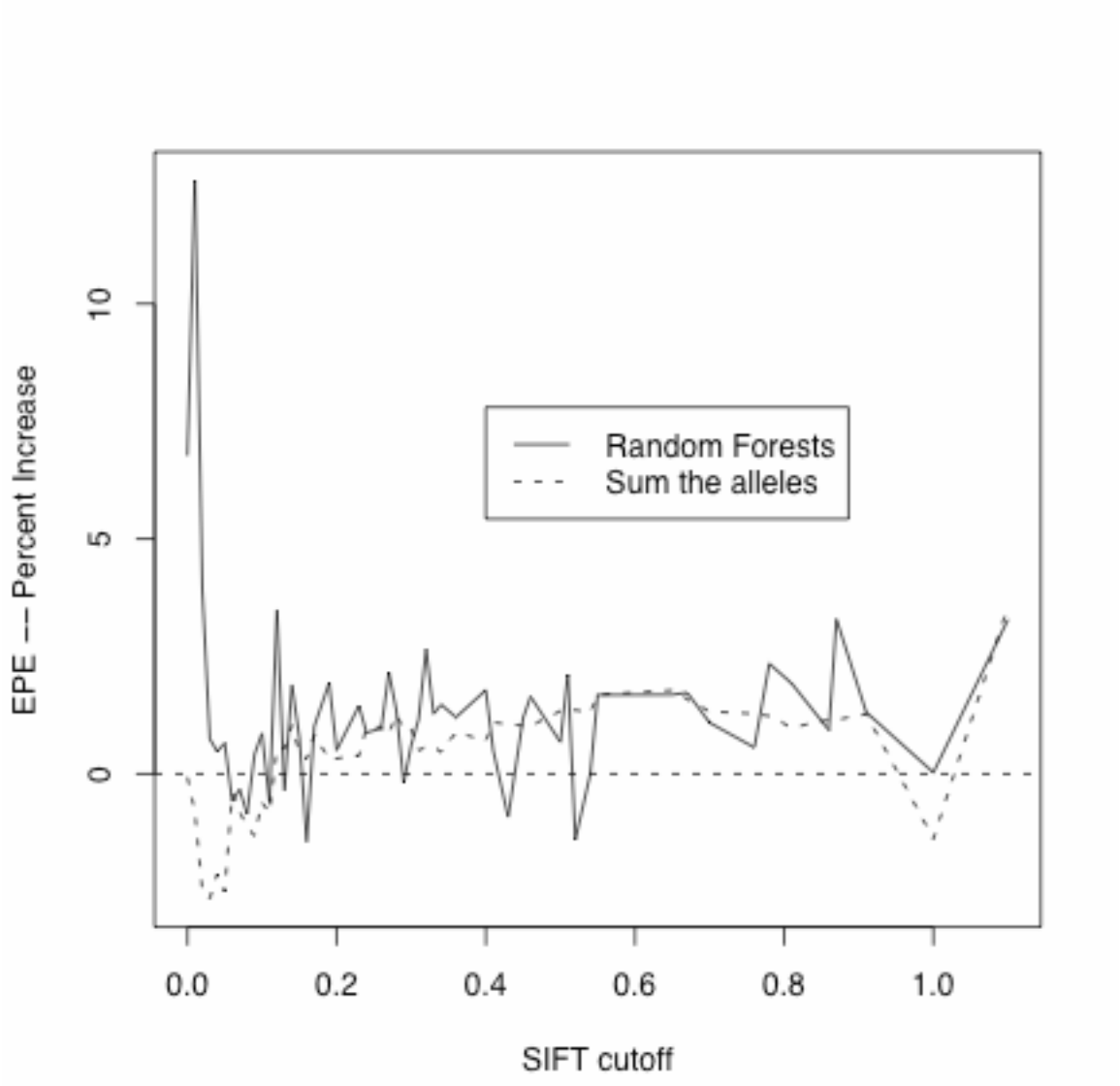

Fig. 4. Expected prediction error for predicting background DNA damage for the best Random Forests Regression model with 1, 2, …, 15 variables. The horizontal line shows the expected prediction error for the null model, i.e., the variance of background DNA damage values as quantified by the Comet Distributed Moment in the alkaline Comet assay. Only the 99 SNPs with at least two variant alleles were used in the analysis.